DOCUMENT RESUME

ED 115 933 CE 005 827

AUTHOR Day, Gerald F.
TITLE An Investigation Into the Use of Criterion-Referenced Measurement in Vocational and Technical Training.
REPORT NO VT-102-309
PUB DATE 75
NOTE 71p.

EDRS PRICE MF-$0.76 HC-$3.32 Plus Postage
DESCRIPTORS Behavioral Objectives; Comparative Analysis; *Criterion Referenced Tests; Educational Research; Feasibility Studies; Literature Reviews; Measurement; Norm Referenced Tests; *State of the Art Reviews; Technical Education; Testing; *Test Reliability; *Vocational Education

ABSTRACT

The paper investigates and analyses the current state of the art of criterion-referenced measurement (CRM), with a view to determining its use in training and instructional programs. It presents a reveiw of the literature pertaining to the following aspects: a brief history of CRM; a definition and comparison of criterion-referenced and norm-referenced measures; usage of the two measures; and the construction and evaluation of criterion-referenced tests in terms of validity, reliability, and other test characteristics. The literature supports the following conclusions: (1) all definitions of CRM stress score interpretation as representing what the individual can do relative to instructional objectives rather than other individuals; (2) criterion-referenced information is valuable in making certain decisions based on what a person can do at a given point in the training cycle; (3) CRM has focused much attention on behavioral objectives and training outcomes; (4) behavioral objectives must be carefully written to effectively direct and measure instruction; (5) more than one measure should be used to validate any CRM to decrease the error associated with its measurement; (6) it is difficult to develop objective procedures necessary for CRM of complex behavior; (7) CRM supplements but should not replace normative tests in training; and (8) more research is needed before extensive use of CRM in instructional programs can be recommended. (Author/NJ)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED115933

AN INVESTIGATION INTO THE USE OF CRITERION-REFERENCED MEASUREMENT IN VOCATIONAL AND TECHNICAL TRAINING

by

Gerald F. Day
Department of Industrial Education
University of Maryland
1975

VT-102-309

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CE 005 827

01 2/3

TABLE OF CONTENTS

Chapter		Page
1.	INTRODUCTION	1
	Purpose and Scope of this Paper	3
2.	USING CRITERION-REFERENCED MEASUREMENT IN TRAINING	5
	Evaluation and Measurement in Training	5
	Historical Perspective of Criterion-Referenced Measurement	9
	Defining Criterion-Referenced Measurement.	12
	Norm-Referenced versus Criterion-Referenced Measurement	14
	Uses for Criterion-Referenced Measurement.	19
	Writing Criterion-Referenced Tests	21
	Evaluation of Criterion-Referenced Tests	30
	Validity	31
	Reliability	34
	Item Analysis	37
	Reporting and Interpretation	40
	The Application of Criterion-Referenced Measurement	42
	Butler's System	43
3.	SUMMARY AND CONCLUSIONS	55
	REFERENCES	60
	APPENDIX A	66

Chapter 1

INTRODUCTION

A great deal of work has been done in training and educational evaluation and measurement since E. L. Thorndike's (1918) declaration of faith, "Anything that exists at all exists in some quantity, and anything that exists in some quantity is capable of being measured." (p. 16)

The concept of criterion-referenced measurement (CRM) has received a great deal of attention recently in training, educational, and measurement literature. Trow (1961) and others have suggested that it may mark the beginning of a new era in measurement. The recent emphasis on CRM has been due to the concern about the measurement of proficiency or competency of occupational and educational tasks.

Glaser (1963), a pioneer in CRM, stated:

> . . . many of us are beginning to recognize that the problems of assessing existing levels of competence and achievement, and the conditions that produce them require some additional consideration. (p. 531)

Glaser (1963) has suggested that what is needed in measuring competency is:

> . . . explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the

> underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of referenced to the performance of others. (p. 520)

A main issue in the CRM movement is the distinction between norm-referenced and criterion-referenced approaches to measurement. Norm-referenced measurement (NRM) indentifies an individual's test performance in relation to the performance of others on the same measure. CRM identifies an individual's performance with respect to specified performance standards.

Jackson (1970) has pointed out that it has become increasingly clear that measurement by norm-referenced tests does not provide the information that is needed in making certain kinds of decisions about instructional programs. Cronbach and Gleser (1965) have questioned the usefulness of classical test theory and NRM for all testing situations. Popham and Husek (1969) concluded that:

> . . . the problem is now not only how to summarize a student's performance on a test, but also how to insure that a test is constructed (and judged) in a manner appropriate for its use, even if its use is not in the classical framework. (p. 1)

Although most of the literature on CRM has come from educational sources, its use has been advocated for industrial, military, business, and governmental training programs and promotions. (Fremer, 1972; Garvin, 1971; Goldstein, 1974; Swezey, Pearlstein, and Ton, 1974; Thronton and Wasdyke, 1972) Goldstein (1974) has pointed out that:

> The norm-referenced measures tell us that one student is more proficient than another, but they do not provide much information about the degree of proficiency in

> relationship to the tasks involved. Unfortunately, many training evaluations have employed norm-referenced measures to the exclusion of other forms of measurement. In order to properly evaluate training programs, it is necessary to obtain criterion-referenced measures that provide information about the skill level of the trainee in relationship to the expected program achievement levels. (pp. 63-64)

Measurement specialists (Cronbach, 1963; Ebel, 1962; Hambleton and Novick, 1973; Livingston, 1972; and Millman, 1974) have indicated that there is a pressing need to develop achievement or performance measurement theory. Glaser and Nitko (1971) have asserted that:

> Tests that measure instructional outcomes and that are used for making instructional decisions demand special characteristics--characteristics that are different from the mental test model that has been successfully applied in aptitude testing work. (p. 652)

Purpose and Scope of This Paper

The purpose of this paper was to investigate and analyze the current state-of-the-art of CRM to determine the feasibility of using it in training and instructional programs.

The following questions were posed by the writer in an attempt to analyze CRM:

1. What is criterion-referenced measurement?
2. What are the differences and similarities between criterion-referenced and norm-referenced measures?
3. When and how should criterion-referenced measurement be used?
4. How is a criterion-referenced test constructed?

5. How can a criterion-referenced test be evaluated in terms of validity, reliability, discrimination, and other test characteristics?

Throughout this paper, training and education has been used interchangeably. It is the belief of the author, and that of others, that education and training deal with the same instructional processes of acquiring skills, knowledges, and attitudes in order for an individual to perform in another environment. As Goldstein (1974) has pointed out, both of the disciplines deal with similar areas, such as specification of objectives,environmental design, and evaluation.

Writings of those in education and those in other fields have tried to be synthesized. However, by the very fact that most of the literature has come from education, this integration was difficult.

A review of the literature in Chaper II pertains to the following aspects: a brief history of CRM; defining the term criterion-referenced measurement; a comparison between CRM and NRM; usage of CRM and NRM; writing a criterion-referenced test; and empirical and logical evaluation of criterion-referenced tests.

The summary and conclusions are included in Chapter III.

Chapter II

USING CRITERION-REFERENCED MEASUREMENT IN TRAINING

Since 1963, the area of CRM has been a hot topic, with hundreds of articles and books being written about its theoretical basis, development, use, and test parameters. This section attempted to analyze and synthesize the current state-of-the-art of CRM and its role in training.

Evaluation and Measurement in Training

In the instructional process, learning has been defined as:

> . . . the process by which behavior is initiated or changed as a result of experience . . . through training and practice. (Garry, 1963, p. 2)

The particular aspects of behavior acquired by an individual depend upon how the training environment is designed and developed. What is taught and how it is taught depends upon the objectives and values of the organization. (Lynton and Pareek, 1967)

Many facets of human behavior are involved in the instructional process: the learning of the subject matter content and skills; and the processes involved in using them, such as critical thinking, retention, transfer, problem solving, and creating. The attitudes and motivation toward these activities are also forms of behavior. The total design of a training environment is a complex enterprise, and there

are many variables which foster, nuture, guide, influence, and control human behavior within its structure. (Lynton and Pareek, 1967)

Evaluation and measurement play an important role in the instructional process. It should be noted, however, that the terms "evaluation" and "measurement" have distinctive meanings. Measurement is concerned with the application of an instrument or instruments to collect data for some specific purpose. (Green, 1970) In other words, measurement refers to quantitive descriptions of behavior, things, or events. (Gronlund, 1968)

Evaluation is a broader concept than measurement in that it involves not only quantitative descriptions, but also qualitative descriptions. Gronlund (1968) wrote:

> In addition to such numerical and verbal descriptions, evaluation includes value judgements concerning the thing described. Thus, when we evaluate the achievement of a student, the effectiveness of instruction, or the appropriateness of a curriculum, we are concerned with judging their value or worth. (Gronlund, 1968)

Evaluation is a comprehensive and complex process. The procedural steps, as described by Gronlund (1968), include:

> . . . (1) identifying the objectives (i.e., the desired outcomes), (2) defining the objectives in behavioral terms (i.e., specifying the behavior we are willing to accept as evidence of the desired learning), (3) selecting, or constructing, instruments for measuring (or describing) the behavior, and (4) applying the instruments and analyzing the results to determine the degree to which the desired learning outcomes have been achieved.

The fundamental task of measurement is to provide information for making, basic, essential decisions with

respect to the instructional design and operation. (Nelson, 1970) According to Glaser and Nitko (1971), four activities of instructional design determine measurement requirements. These are :

> . . . the analysis of the subject-matter domain under consideration, diagnosis of the characteristics of the learner, design of the instructional environment, and the evaluation of the learning outcomes. (pp. 625-626)

In the analysis of the subject matter, experts analyze the subject matter domain in terms of performance competencies. The characteristics of the domain are constructed according to conceptual hierarchies and operating rules in terms of increasing complexity of human performance. The analysis and definition of instructionally relevant performance is of major concern. This can be accomplished through the specification of behavioral objectives, translating them into types of observable performance, and conducting research studies about different instructional methodologies. (Glaser and Nitko, 1971)

Diagnosing the characteristics of the trainee involves the measurement of the behavior an individual has upon entering a program. In other words, these measurements provide information about existing pre-instructional behavior. This is helpful in starting the instruction based on what the trainee already knows and can do. (Goldstein, 1974; Millman, 1972; Mirsberger, 1974)

The third activity is that of designing the instructional environment and specifying the conditions under which

learning can take place. This allows the individual to progress toward the training goals described as subject-matter competence and acquire the desired outcomes of instruction. (Glaser and Nitko, 1971)

The final activity of evaluation is measuring learning outcomes. This provides information about the extent to which the instructional objectives have been attained and the extent to which the behavior of the trainee approaches the performance criteria. The trainee is said to have mastery of the instructional objectives when the degree of performance has been attained as specified by the designers of the instructional program. (Glaser and Nitko, 1971)

Mirsberger (1974) stated that:

> Evaluation, in the view of the trainee-oriented instructor, is the process of obtaining feedback which is then used to direct the remaining portion of the training program. (p. 34)

Mirsberger's phases are similar to Glaser and Nitko's (1971) stages, but he adds an on-the-job performance phase. Mirsberger's phases include:

1. Pretraining phase: that evaluation done before any actual training is started.
2. Training phase: evaluation made throughout the learning period.
3. Posttraining phase: the evaluation made at the end of the training effort.
4. Performance phase: the evaluation of the matriculated trainee in an on-the-job situation after the training effort. (p. 34)

In summary, learning in a training environment is a process of changing the behavior of an individual from an initial entering state to a specified terminal state.

Instruction is the practice of providing conditions and activities for this transaction to occur. Evaluation, of which measurement is a part, is the collecting of data, assessments, and information about the instructional program and the trainee's performance. It is used to make basic decisions in developing the overall effectiveness of the training system.

Historical Perspective of Criterion-Referenced Measurement

The psychological testing movement started with the Darwinian emphasis on differences between individuals, and the theorical framework of test scores was developed to emphasize differences in abilities and traits. (Mehrens and Lehmann, 1969) Psychological testing has concentrated on comparative interpretations. What the mental test measures is whatever causes some people to get high scores, and others to get low scores. The psychologist is likely to say that the test measures nothing if everyone scores the same, except for variation due to errors of observation. (Cronbach, 1971)

With the development of psychological tests around the turn of the Twentieth Century by Galton, Cattell, Binet, Goddard, Terman, Otis, and others, a new era in measurement was born. The mental test (a term coined by Cattell in 1890), although developed to discover and predict aptitude, was introduced in the schools to measure achievement for diagnostic and training purposes. (Trow, 1971)

Achievement testing is different from aptitude

testing in that:

> An achievement test is used to measure an individual's present level of knowledge or skills or performance, an aptitude test is used to predict how well an individual may learn. (Mehrens and Lehmann, 1969, p. 73)

After World War I, there was a boom in standardized subject-matter tests, statistics and measurement courses, and textbooks related to these fields. (Horrocks and Schoonover, 1968)

Although the mental test was devised to differentiate and compare individuals for recommending further treatment, training, or education, the procedures of assigning school marks got mixed-up in their use. Because the system of assigning grades was based on the probability curve, the mark a student received was based on what others did on the same test,not on what level of knowledge, understanding, or skill proficiency the individual pupil had achieved. (Trow, 1971, p. ix)

Recently, there has been a revival of interest in absolute measurement, now retitled criterion-referenced measurement. (Ebel, 1962; Glaser, 1963; Popham and Husek, 1969; Tyler, 1966) CRM has been around in this country since the early part of the Twentieth Century, with scales developed by Courtis, Thorndike, Ayres, and others for measuring handwriting, compostion,arithmetric and other subjects. (Trow, 1971) During the period from 1909 to 1915, a series of arithmetic tests and five scales for measuring abilities in English composition, spelling, drawing, and handwriting

were published. (Odell, 1930)

In 1909, Thorndike publihsed a standardized achievement scale, The Scale for Handwriting of Children. The introduction of standard measures of achievement is most often attributed to E. L. Thorndike, whose students were later to make great contributions to the field of measurement and achievement testing. (Horrocks and Schoonover, 1968)

Ayres' handwriting scale was devised by judges who studied and arranged different specimens of pupil handwriting according to quality. Suitably spaced specimens were selected to represent different levels of proficiency and these were reproduced as a guide for teachers. A teacher could simply look at successive Ayres' scores on a pupil's cumulative record and judge how the pupil's handwriting was progressing. (Cronbach, 1971)

Ebel (1965) has pointed out that the percentage-mastery grades which were once widely favored in schools in the early 1900's represented a crude type of criterion measurement, although one that was generally unsatisfactory in practice.

In 1913, Thorndike noted the limitations of NRM and grades since they did not indicate the mastery, amount, or type of skills and knowledges possessed by the student. Thorndike (1913), in discussing the assigning of school grades based on normative data, stated:

> . . . the vices of the old system . . . was its relativity and indefiniteness--the fact already described that a

> given mark did not mean any defined amount of knowledge, or power, or skill--so that it was bound to be used for relative achievement only.
>
> The proper remedy is not to eliminate all stimulus to rivalry, and along with it a large part of the stimulus to achievement in general, but to redirect the rivalry into tendencies to go higher on an objective scale for absolute achievement, to surpass one's own past performance, to get into what, in athletic parlance, is called a 'higher class,' to compete within that class, and to compete cooperatively as one of a group in rivalry with another group. (pp. 287-288)

Nevertheless, the old NRM system which Thorndike referred to is the one that is still used today by the majority of evaluation experts. (Trow, 1971)

Defining Criterion-Referenced Measurement

Glaser has been credited with having introduced the current-day definition of CRM. (Jackson, 1970) In one of Glaser's more recent writings on the subject, the following definition was suggested:

> A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. (Glaser and Nitko, 1971, p. 653)

Glaser (1963) stated that criterion-referenced tests can be differentiated from norm-referenced tests in that they do not focus on the problem of individual differences. Rather, they are aimed at indicating what an individual can do and cannot do.

Although Glaser's definition is the classical one used by most people, it is not the only one. Popham and Husek (1969) have proposed:

Criterion-referenced measures are those which are

> used to ascertain an individual's status with respect to some criterion: i. e., performance standard. It is because the individual is compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced. The meaningfulness of an individual's score is not dependent on comparison with testees. We want to know what the individual can do, not how he stands in comparison with others. (p. 2)

Ebel (1971) characterized CRM in terms of score distribution and interpretation:

> The essential difference between norm-referenced and criterion-referenced measurements is in the quantitive scales used to express how much the individual can do. In norm-referenced measurement the scale is usually anchored in the middle, on some average level of performance for a particular group of individuals. The units on the scale are usually a function of the distribution of performances above and below the average level. In criterion-referenced measurement the scale is usually anchored at the extremities, a score at the top of the scale indicating complete or perfect mastery of some defined abilities, one at the bottom indicating complete absence of these abilities. The scale units consist of subdivisions of these total score ranges. (p. 282)

Wang (1969) has expressed that a criterion-referenced test ". . . is an achievement test developed to assess the presence or absence of a specified criterion behavior described in an instructional objective." (p. 14)

It is interesting to note that these various definitions agree in that they emphasize the <u>direct interpretability</u> of scores, but differ in the extent to which they make reference to the method by which the test is constructed. Ebel emphasized the scale from which interpretations are to be made, while Glaser stressed the construction.

Most writers stress the method of construction, such as Jackson (1970) who wrote:

> . . . the term 'criterion-referenced' will be used here to apply to a test designed and constructed in a manner that defines explicit rules linking patterns of test performance to behavioral referents. (p. 3)

The preceding concepts are somewhat different than one other prevalent use of the term criterion-referenced used in psychometric literature. That principle involves correlating the scores of an achievement measuring instrument (X) with a second measurement situation (Y), such as another test or grade average. The Y score would be referred to as a criterion score and the degree of relationship is expressed by the product-moment correlation. (Tuckerman, 1972) Criterion-related validity is similar to this concept in that it is a technique for showing the relationship between test scores and an independent external measure, such as a standardized test. (Karmel, 1970)

Norm-Referenced versus Criterion-Referenced Measurement

The heart of the issue concerning CRM and NRM is deriving meaning from the test score. The score received by an individual on any type of test is basically inert and and must be related semantically to the behavior of the individual. (Lord and Novick, 1968) Ebel (1962) stressed that:

> No test score, raw or standard, has much meaning as an abstract number. Additional data for interpretation must always be provided, either by the test producer or by the test user from his own knowledge and experience. The numbers which report standard scores are no more intrinsically meaningful, and no more self-interpreting, than raw scores. (p. 16)

For the most part, measurement specialists have concentrated on interpreting the test scores primarily based on the scores of others. At present, the use of NRM is almost universal in the United States. (Ebel, 1962)

Advocates of CRM are trying to operationally define standards upon which interpretations can be made directly from the score. These experts believe that norm-referenced interpretations have serious limitations ". . . when they are employed with achievement tests that are used in instructional systems seeking to be adaptive to the individual." (Glaser and Nitko, 1971, p. 653)

According to Glaser and Nitko (1971), NRM has been so dominant in training and education because of the:

> . . . concentration of psychological test theory on trait variability and on the relative difference between individuals, the reluctance of educators to specify precisely their goals in terms of observable behavior, the reliance of measurement specialists on the mental test model, and the desire of test constructors to build tests that are applicable to many different instructional systems. (p. 657)

As Popham and Husek (1969) have observed, it is impossible to tell a norm-referenced test from a criterion-referenced test by just looking at it. The difference is found by examining the purpose of the test, the manner in which it was constructed, the specificity of the information obtained about the domain of instructionally relevant tasks, the generalizability of the test performance, and the use of the scores.

Arguments have been made that any achievement test

defines a criterion because it is representative of desired outcomes, and that one can determine the particular tasks an individual can perform by just examining the responses on the person's test. Jackson (1970) wrote:

> Any test samples the content of some specified domain. Even though a test may be normed so that an individual's score may be compared with scores of some specified group, there is the assumption of some latent trait upon which observed scores depend, and which the test is, therefore, said to measure. Hence, there is always an implicit behavioral element, and even tests that are described as norm-referenced are designed to yield inferences about, say, the amount of trait X that an individual has. In contrast to a criterion-referenced test, however, the inference is of the form--more (or less) of trait X than the mean amount in population Y--rather than some specified amount that is meaningful in isolation. (p. 2)

However, Glaser has argued that the way a normative test is constructed and designed negates its use as a true criterion based on performance standards. In practice, desired outcomes have seldom been specified in performance terms prior to constructing a norm-referenced test. (Glaser and Nitko, 1971) When using a NRM, questions that appear on the final criterion test have been revised and arranged to maximize the test constructor's concept of what the distribution of final scores should be and how the terms should function statistically. (Cox, 1971)

Other determinates of test construction have been ease of administration and scoring. Lindquist (1968) has indicated that many valuable instructionally relevant tasks are not being tested because of computer-scoring restrictions.

All of these practices tend to distort the results of a

person's score with respect to a clearly defined domain of tasks and performance standards. (Glaser and Nitko, 1971)

With respect to specificity of the information obtained by CRM about the domain of tasks, there should be a logical transition from the domain to the test and vice versa. There should be little difficulty in identifying the class of tasks that can be performed. Thus, all tasks in the domain must be defined in observable behavior. (Thornton and Wasdyke, 1972)

The attainment of certain abilities, skills, and knowledges can only be inferred based on observable performance. In an occupational area, the specified domain of tasks would be analyzed and broken down into observable performance measurement. Criterion-referenced tests do not seek to indicate how much ability a student possesses along a hypothetical ability dimension, but whether certain kinds of tasks can be demonstrated. This implies an analysis of task structure in which each task description includes criteria of performance. In turn, a scoring system must be devised that will preserve information about the tasks that an individual can perform. (Fremer, 1972) Norm-referenced scores, such as percentile ranks, t-scores, and grade equivalents lose the specificity of criterion information. (Ebel, 1962)

There must be generalizability of test performance to total task domain. As the trainee progressess in a

program, the number of tasks become very large. The criterion-referenced test constructor must determine how long to make a test so that generalization can be made about which specific tasks a learner can perform. The norm-referenced test constructor does not have this problem since wide selection of items will result in variable scores so that it can be said that individual X can do more or has achieved more than individual Y. However, what individual X can actually do is really not known. An individual's item responses provide only a weak basis for inference when norm-referenced tests are used.

Table 1 shows key features of CRM and NRM, as interpreted by Boehm (1973).

	Norm-Referenced	*Criterion-Referenced*
1. General Purpose	To make comparisons among individuals	To determine *how* an individual functions relative to a criterion
	To make decisions about placement in programs in which only limited numbers of individuals can be accepted	To program specifically for the individual
	To determine for whom a program "works"	To determine whether an instructional program "works" in developing criterion behaviors
2. Item Types	Items must discriminate among individuals	Items must correspond to criterion levels
	Items all subjects pass or all fail eliminated	Items must provide explicit information about *what* an individual can or cannot do
3. Content	Content may or may not match particular classroom goals	Content *must match* classroom objectives which have been behaviorally defined beforehand
	Sampling is made from the larger task domain	Criterion levels can be set at each content level of a program and must specify minimal levels of competence
4. Scores	Variability among scores is essential	Variability is irrelevant
	Scores can mask what an individual can do but provide indication of his relative standing	Scores must reflect (not mask) what an individual can or cannot do
5. Type of Ranking	Use of age and grade norms; percentiles; standard scores	Percentage passing a criterion level
		Pass/fail information on each item

Table 1. Characteristics of Norm-Referenced and Criterion-Referenced Tests (Boehm, 1973)

Uses for Criterion-Referenced Measurement

An important consideration in deciding which type of measurement to use is the use of the scores. Although both CRM and NRM provide data for decision making about individuals and treatments, the context with which decisions are made determine which to use.

NRM should be used if there is some degree of selectivity necessary, such as a limitation to the number of people that can be admitted to a training program. (Popham and Husek, 1969)

CRM should be used to make decisions about individuals and treatments in other situations. A criterion measure could be used to determine whether a person has mastered certain skills considered a prerequisite to starting a new training program. A criterion measure reflecting a set of instructional objectives could be used to evaluate two different instructional sequences to determine which is more effective. If competencies possessed by an individual is needed before instruction can be provided, CRM should be used. (Popham and Husek, 1969)

Other suggestions have been made for using CRM. Coulson and Cogswell (1965) discussed the need for it in regard to the use of programmed materials. Glaser and Cox (1968) suggested the use of it in individualized instructional models where evaluation instruments must differentiate between groups of pupils who have mastered certain units

and those who have not. Jackson (1970) concluded that CRM would be desirable in the areas of diagnostic information, formative evaluation of training programs, and the evaluative assessment of individual and group achievement. Fremer (1972) suggested that CRM is meaningful in relating performance to significant real-life criteria such as minimal competency in a basic skills area, such as math for an accountant. Thronton and Wasdyke (1972) advocated its use in performance-based evaluation for job promotions and certification, such as in "The New York City Police Study for Promotions" and the National Teacher Examination in Industrial Arts.

Garvin (1971) has suggested that different levels of proficiency standards be established for certain occupational tasks. If certain tasks, by their very nature, must be performed at a specifiably high level, than an absolute criterion level should be established and met by all. For example, landing an aircraft or compounding a prescription must be done correctly or public safety would be endangered. However, there are other tasks where some latitude of competence is permissible, such as running a lathe, selling a product, and typing. Different levels of proficiency could be established for these relative tasks.

Garvin (1971) further set forth some general principles regarding the applicability of CRM to various content areas and levels:

1. Unless at least one of the instructional objectives of a unit envisions a task that must be subsequently

be performed at a specified level of competence in at least some situation, criterion-referenced measurement is irrelevant because there is no criterion. In this sense the entire sequence of 'social studies' provides no meaningful criterion except, possibly, the entry level for certain 'honor' courses.

2. If public safety, economic responsibility, or other ethical considerations demand that certain tasks be performed only by those 'qualified' for them by formal instruction, then CRM of the outcomes of such instruction is clearly indicated. The criterion here is the licensing standards of the profession involved. All professional instruction in the medical arts, law, finance, engineering, and the applied physical and social sciences generally is clearly in this category. Teaching--at any level--ought to be. However, entry to such professional training is typically based on NRM since training capacity imposes a 'quota.'

3. In any instructional sequence where the content is inherently cumulative and the rigor progressively greater, CRM should be used to control entry to successive units. However, if there are several different sequences differing widely in rigor, NRM is more useful in making appropriate placements.

4. There are certain content areas to which criteria do apply but not everyone need meet then. These are the 'required subjects', everyone must try to learn them--if only as a matter of public policy--but it is almost preordained that some of them will not. Home economics and physical education are relatively non-controversial examples at the secondary level; at the college level, these become professions and CRM applies. (pp. 62-63)

Most test experts stress, however, that both criterion-referenced and norm-referenced measures are needed to make valid and enlightened decisions about individuals and programs. (Simon, 1969; Swezey, Pearlstein, and Ton, 1974)

Writing Criterion-Referenced Tests

The areas of writing CRM's and evaluating criterion-referenced tests are in the developmental stage. Many people have written articles hypothesizing how to write a criterion-referenced test and evaluate it in terms of validity,

reliability, and other test parameters. However, there is need for developing a CRM test theory. (Boehm, 1973; Glaser and Nitko, 1971; Hively, 1974; Jackson, 1970) In a 1974 poll of its members, the National Council on Measurement in Education found that the development of a test theory for CRM was ranked number three in its priority list for research in measurement. The following two sub-sections discuss various writings in the field.

An important concept to be cognizant of when writing a CRM, is that of a criterion. Although most writers do not emphasize the theoretical basis for criteria, Goldstein (1974) has pointed out that criterion relevancy, deficiency, and contamination are important concepts to be aware of. Nagle (1953) stated that a criterion is more relevant when the criterion measure is closer to the true criterion. Thorndike (1949) emphasized that the criteria are more relevant if the behaviors learned in the training program are the same as those required for success at the ultimate task. (Goldstein, 1974)

Since Travers (1975) has covered behavioral objectives, it is sufficient to say here that after the organizational needs assessment and task analysis, behavioral objectives should be written. Most CRM people have used Mager's (1962) format. These objectives must be translated into specific test tasks that form the basis for inference that the behavior has been acquired by the trainee if successfully completed.

Recently, much work has been done in the analysis and classification of behavior in training and education, and

this has been helpful in analyzing performance into component tasks. (Bruner, 1964; Gagne, 1965; Glaser, 1962; Melton, 1964; Miller, 1965) Other studies (Gane and Woolfenden, 1968; Gibson, 1965; Hively, 1966; Newell and Forehand, 1968) have dealt with examining the specific components and the sequence of performance of a complex behavior so that the task domain can be identified for training and testing purposes.

Specifying the domain of tasks requires a systematic procedure. Hively (1968) has developed one method to delimit and clearly define the domain of tasks through the use of an "item form." Table 2 contains examples of item forms for subtraction tasks in arithmetic. A title in the left column contains a task of the subtraction domain. Next, a sample problem is shown as it would appear on the test. The last two columns contain the general form and generation rules which define the tasks. A collection of item forms constitute a domain from which test items may be drawn. Using item forms, it is easy to make judgements about the content validity of a criterion-referenced test, or in fact, any kind of test.

Osburn (1968), who has developed a similar item form, discussed two conditions that are prerequisites for allowing inferences to be made about a domain of skills and knowledge from performance on a sample of items:

> The first is that all items that could possibly appear in the test should be specified in advance. Secondly, the items in a particular test should be selected by random sampling or stratified random

sampling from the universe of content. (p. 96)

Descriptive Title	Sample Item	General Form	Generation Rules*
Basic fact; minuend > 10.	13 −6	A −B	1. $A = 1a_1$; $B = b$ 2. $(a < b) \in U$ 3. $\lvert H, V \rvert$
Simple borrow; one-digit subtrahend.	53 −7	A −B	1. $A = a_1a_2$; $B = b$ 2. $a_1 \in U - \lvert 1 \rvert$ 3. $(b > a_2) \in U_0$
Borrow across 0.	403 −138	A −B	1. $N \in \lvert 3, 4 \rvert$ 2. $A = a_1a_2 \cdots$; $B = b_1b_2 \cdots$ 3. $(a_1 > b_1)$, $(a_3 < b_3)$, $(a_4 \geq b_4) \in U_0$ 4. $b_2 \in U_0$ 5. $a_2 = 0$ 6. $P\lvert\lvert 1, 2, 3\rvert, \lvert 4\rvert\rvert$
Equation; missing subtrahend.	42− =25	A− =B	1. $A = a_1a_2$; $B = b_1b_2$ 2. $a_1 \in U$ 3. $a_2, b_1, b_2 \in U_0$ 4. Check: $0 < B < A$

* Explanation of notation:

Capital letters $A, B, \cdots$ represent numerals.
Small letters (with or without subscripts), a, b, a_1, b_2, etc., represent digits.
$x \in \lvert \cdots \rvert$: Choose at random a replacement for x from the given set.
$a, b, c, \in \lvert \cdots \rvert$: All of a, b, c are chosen from the given set *with replacement*.
N_A: Number of digits in numeral A.
N: Number of digits in each numeral in the problem.
$a_1, a_2, \cdots \in \lvert \cdots \rvert$: Generate all the a_i necessary. In general "$\cdots$" means continue the pattern established.
$(a < b) \in \lvert \cdots \rvert$: Choose two numbers at random *without* replacement; let a be the smaller.
$\lvert H, V \rvert$: Choose a horizontal or vertical format.
$P\lvert A, B, \cdots \rvert$: Choose a permutation of the elements in the set. (If the set consists of subscripts, permute those subscripted elements.)
Set operations are used as normally defined. Note that $A - B = A \cap \bar{B}$. Ordered pairs are also used as usual.
Check: If a check is not fulfilled, regenerate all elements involved in the *check* statement (and any elements dependent upon them).

Special sets:

$U = \lvert 1, 2, \cdots, 9 \rvert$
$U_0 = \lvert 0, 1, \cdots, 9 \rvert$

Table 2. Examples of item forms from the subtraction universe developed by Hively. (Hively, 1968)

Jackson (1970) stated that ". . . the difficulty of objectively defining a test construction process is directly related to the complexity of the behavior the test is designed to assess." (p. 7) Thus, the first of Osburn's conditions would be difficult to satisfy for complex domains. However, listing the elements of a universe of item content can be

overcome, to a certain extent, if a generative process could be defined which could, in theory, produce such a listing. Through the use of the item form, it is possible to produce such a generative process. (Hively, 1968; Osburn, 1968)

Osburn (1968) has described the characteristics of an item form as follows:

> . . . (1) it generates items with a fixed syntactical structure; (2) it contains one or more variable structures; and (3) it defines a class of item sentences by specifying the replacement sets for the variable elements. (p.96)

Using the item form method, there is an "unbroken link" between the generative system and the specific item produced. A collection of item forms, together with the replacement sets for the variable elements, then define a universe of content. In addition to the numerical type of Hively's, Osburn has developed verbal replacement sets and a hierarchical arrangement of test tasks to be generated.

An item form could consist of a sentence with one or more blanks, and the words or numbers that fit into the blanks could be systematically varied to produce items of different levels of specificity. Since this procedure is systematic and rule bound, it has been adaptable to computer programming. (Ferguson, 1969) Shoemaker and Osburn (1969) have constructed a computer program ". . . capable of generating random or stratified random parallel tests from a specified content population." (p. 165)

An example of a sentence frame for the input of a computer program would be:

> Given a normal distribution with a mean equal to____ and a standard deviation equal to ____. If one number is randomly sampled from this distribution, what is the probability that this number will be greater than or equal to ____? (Shoemaker and Osburn, 1968)

The blanks in the form are filled in by a random number generator, which can be controlled to supply realistic problems and reduce difficult and long computations. (Shoemaker and Osburn, 1969)

Bormuth (1970) has advocated that the tests that use NRM procedures cannot unequivically claim to represent the properties of instruction nor can they be objectively reproduced. A norm-referenced test item, Bormuth wrote, is a property of the test writer and not a property of instruction. A score on a norm-referenced test is the learner's responses to the writer's responses to instruction, or, in other words, the constructor's behavior.

Ebel (1962) reaffirmed Bormuth's beliefs.

> Specialists in educational measurement generally recognize that most objective tests rest on highly subjective foundations. The abilities, values, and idiosyncrasies of the test constructor have played a major part in determining the content of most tests. Test specifications sometimes exist only in the mind of the test constructor or in a few brief written guidelines. When written, they often have more to say about the form of the test than about its content. (p. 22)

Bormuth (1970) has suggested that a linguistic analysis be used to explicitedly translate instructional objectives into test items. Like the item form, this would introduce more objectivity and replicability into test writing.

As Swezey, Pearlstein, and Ton (1974) have shown, there are many studies going on with CRM in different areas. One of the more extensive studies on criterion-referenced testing was done by Thornton and Wasdyke (1972). These test specialists have developed "The Taxonomy of Behavior for Career Development and Measurement" which provides a framework for the logical tracing of observed behaviors from the processes of job analysis, through test development, performance evaluation to validation. The taxonomy can be used to write comprehensive test specifications for simple to complex ranges of behaviors.

There are five steps in Thornton and Wasdyke's (1972) method:

> 1. Job (task) analysis and specification (in task analysis statements).
> 2. Translation and classification of task analysis statements into behavioral objectives.
> 3. Definition of the job performance standards into behavioral terms.
> 4. Multi-dimensional test specification and development.
> 5. Measurement of performance--validity (translation of occupational test items into behavioral objectives). (p. 3)

The above procedure was used for an examination constructed by the Education Testing Service for police promotion procedures in New York City.

The first step in the above process results in an ordered collection of task statements which describe the duties and responsibilities of a job. The second step translates task statements into behavioral objectives, indicating

the condition, performance, and extent. (This is very similar to Mager's procedure for writing behavioral objectives.) This results in a list of behavioral objectives required for acceptable job performance. Each objective is then described in terms of the cognitive activity, the affective mode necessary, and the psychomotor skills required for satisfactory job performance. After this, each objective is classified within a three-dimensional, 90-cell taxonomy of behavior and this serves as a blueprint of the terminal objectives of the process selection (prediction), training (education and career development), and evaluation (performance). (Thornton and Wasdyke, 1972)

In defining job performance standards, judges are used to determine what precisely is the minimal acceptable job performance in terms of that behavior. The results of this process are twofold: minimum acceptable behavior for developing a test for minimum competency; and, the precise lower limit of acceptable job performance specified in a behavioral scale. (Thornton and Wasdyke, 1972)

The final behavioral objectives can be used to write multi-dimensional test specifications. The specifications include the behavior to be measured in the test and the precise level or levels within the taxonomy which most appropriately measures the required job behavior. (Thornton and Wasdyke, 1972)

The last step of measuring performance and validation

is logically determined in two ways:

> 1. The translation of the test items into behavioral objectives, their classification by means of the taxonomy, and their comparison, objective by objective, with the original task derived taxonomy.
>
> These operations are performed by researchers other than the job analysts and test designers.
>
> 2. The comparison of candidate's performance, behavior by behavior, on the test and as rated by supervisors on the job. (Thornton and Wasdyke, 1972, p. 12)

An example of translating a test item into a behavioral objective would be:

> Condition: Given witnesses to a crime in a physical situation in which they cannot be separated from each other,
>
> Performance: Predict the effect of this situation on the information gathered from these witnesses in two areas, the sequence of the events and the description of the perpetrator.
>
> Extent: Accuracy of prediction of 100% based on correct answer in each case. (Thornton and Wasdyke, 1972, p. 12)

This item objective would be traced through the taxonomy back to the original objective and accepted performance standard.

Thornton and Wasdyke (1972) have expressed that this logical validation is not a substitute for statistical validity, but a supplement to traditional methods.

These methods are some of the recent developments in criterion-referenced test construction. The major goal of all of these methods is to be able to allow inference from test performance to behavioral referents. All items are specified by rules and there is the advantage of being able to randomly sample items from a specified universe of content. Work is being carried on by several universities and test services

to refine these methods. (CTB/ McGraw Hill; Educational Testing Services; Army Research Institute)

Evaluation of Criterion-Referenced Tests

After defining the universe of content and constructing the item forms, the final form of the test must be constructed. Item selection and analysis have been well-developed for NRM but not for CRM. While NRM depend on variance in the test scores, CRM may display very little variance. (Popham and Husek, 1969)

For example, if a training program for sewing machine operators seeks to reach a certain level of competence, a pretest-posttest experimental design could be used. Scores on the posttest should show an increased mean performance and a decrease variance since all trainees are expected to acquire knowledge and skill mastery of sewing concepts. (Popham and Husek, 1969)

It should be noted at this point, however, that using CRM's do not limit achievement or competency beyond a certain performance level. As Glaser and Nitko (1971) have stated:

> In theory, adaptive instruction seeks to ensure that all individuals in the population show certain levels of mastery in the instructional domain, while not excluding differences in achievement beyond the general level of mastery established. (p. 659)

Concerning the evaluation of CRM's, measurement specialists cast doubt on applying the conventional empirical evaluation procedures of the mental test theory for assessing reliability, validity, and analyzing test items. With NRM's,

the more variability, the better since the purpose of the test is to spread individuals out. However, with CRM's, variability is irrelevant. The meaning of the score flows directly from the connection between the items and the criterion. (Cox, 1971)

The subtle implication of this central difference is that all traditional theories and formulas for determining what a "good" test is can no longer be used with criterion measures. Most of the formulas for test adequacy indices rely on the concept of variability. (Popham and Husek, 1969)

Specialists have stresses that a criterion-referenced test may be a good test even if there is no variance in the population's scores. Indeed, with some criterion tests, it may be that all students will pass every item ! (Cartier, 1968)

Validity. Tuckerman (1972) has defined validity of a test as ". . . the extent to which a test measures what it purports to measure." (p. 139) For example, a test on repairing automobile ignitions must be a true indication of a student's skill and knowledge of automobile ignitions, and not mathematics or reading.

Validity, which is essential for any good test, has been defined in many ways throughout the years. Gulliksen (1950) has stated, "The validity of a test is the correlation of the test with some criterion." (p. 68) Cureton (1951) wrote, "The validity of a test is an estimate of the correla-

tion between the raw test scores and the 'true' criterion scores." (p. 625) Lindquist (1942) has defined validity as ". . . the accuracy with which it measures that which it is intended to measure, or as the degree to which it approches infallibility in measuring what it purports to measure." (p. 213) Edgerton (1949) has stated, "By 'validity' we refer to the extent to which the measuring device is useful for a given purpose. (p. 52) Cronbach (1960) has advocated, "The more fully and confidently a test can be interpreted, the greater its validity." (p. 1151)

There is a conceptual similarity between these statements, but there is also some distinctive differences. The first two deal with correlations, the third avoids statistics, the fourth stresses utility, and the fifth relates to interpretability of the test scores.

The American Psychological Association has identified three basic types of test validity. Content validity is the extent to which as test measures a representative sample of the subject matter content and the behavioral change under consideration. Criterion-related validity is the extent to which test performance is related to some other valid measure. Construct validity is the extent to which test performance can be interpreted in terms of certain psychological constructs. (Gronlund, 1971, pp. 78-90)

The last two procedures for assessing validity are based on correlation and thus variability. Hence, they would

not be too accurate for CRM's. Content validity, by its very nature, is the most suitable to validate a criterion-referenced test. (Swezey, Pearlstein, and Ton, 1974)

Content validity is best evidenced by comparing the test content to the universe of content and behaviors being measured. Mehrens and Lehmann (1969) stated that this is accomplished by:

> . . . a comparison of the test content with courses of study, instructional materials and statements of instructional goals, and by critical analysis of the processes required in responding to the items. (p. 310)

Test experts have used different methods to verify content validity. Popham and Husek (1969) have suggested that the general procedure for validating a CRM would be judgement ". . . based on the test's apparent relevance to the behaviors legimately inferable from those delimited by the criterion." (p. 6) Osburn (1968) stressed that a CRM must have content validity built into it because:

> What the test is measuring is operationally defined by the universe of content as embodied in the item generating rules. No recourse to response-inferred concepts such as construct validity, predictive validity, underlying factor structure or latent variable is necessary to answer this vital question. (p. 97)

Content validity can also be determined by using Hively's (1968) item form, which consists of a complete set of rules for generating a domain of test items for a specific objective. Independent experts or judges are used to decide whether or not a test item is congruent with the highly specific behavior domain explicated by the item form.

Thornton and Wasdyke's (1972) method of validation, by rewriting a test item into a behavioral objective and tracing it back through a taxonomy to the original objective, is another way to check validity. Fremer (1972), who has suggested a variety of methods using a panel of judges to validate a test, summed up the feeling of most measurement people by stating, "More than one method should be used to validate any desired criterion-referenced inference." (p. 28)

Reliability. Like validity, there are many ways to describe the reliability of a test. One general definition of reliability is ". . . the extent to which a test is consistent in measuring whatever it does measure." (Mehrens and Lehmann, 1969, p. 368)

Since most of the methods for estimating reliability are dependent upon variance, they cannot be used for CRM's with complete confidence. For example, one of the most common ways to determine internal consistency is by using the Kuder-Richardson formula which relies on score variance. (Tuckman, 1972) However, if everyone on a CRM obtains a perfect score, the internal consistency estimate would be zero, which indicates poor reliability. CRM advocates state that such a test should not be assumed to be poor. In fact, it is possible for a CRM to have a poor internal index and still be a good measure. (Husek and Sirotnik, 1968)

Other typical indices, such as the split-halves method, are also not appropriate for an internal consistency

estimator.

Concerning external consistency estimates, these are also cloudy when used with CRM's. Reliability can be measured by giving the same people the same test on more than one occasion and then comparing each person's performance on both testings. (Tuckman, 1972) However, this test-retest correlation coefficient, dependent on variability, cannot be used either. Popham and Husek (1969) have said that a high inter-item correlation and test-retest correlation is fine and these indices can be used to support the consistency of the test. However, a criterion measure could be highly consistent and yet indices dependent on variability might not reflect that consistency.

Jackson (1970) has proposed a comparison of the scores on two forms of a CRM measuring the same material since criterion-referenced tests should be able to be generated independently and objectively. An index of agreement between the two forms could then be used.

Cox and Graham (1971) have illustrated another way reliability might be viewed using a sequentially scaled test. Theoretically, the test is constructed so that the student answers all items up to his level of attainment and misses all items beyond this certain point. The test uses a Guttman scale, the total score indicating the individual's response pattern. A coefficient of reproducibility is found that indicated how well an individual's response pattern could be

reproduced from knowledge of this total score. This coefficient might be used as a type of reliability estimate across all individuals taking the test.

Livingston (1972a) has proposed a controversial classical test theory approach to CRM, whereby the psych metric theory of true and error scores are used to find the reliability. Livingston has stated that when using CRM, one wants to find out how far a score deviates from a fixed standard. Thus, he has suggested using deviations from a criterion score instead of a mean score (as in NRM), and defines CRM reliability as a ratio of mean squared deviation from the criterion score.

Oakland (1972), Harris (1972), Meredith and Sabers (1972), and others have taken issue with Livingston's model. (For a discussion, see Swezey, Pearlstein, and Ton, 1974)

Swaminathan, Hambleton, and Algina (1974) have proposed that citerion-referenced reliability be defined as:

> . . . a measure of agreement over and above that which can be expected by chance between the decisions made about examinee mastery states in repeated test administrations for each objective measured by the criterion-referenced test. (p. 263)

These specialists believe that the primary purpose of CRM is to classify individuals into mastery categories on the objectives covered by the test. They emphasized that using their method will result in as many reliabilities as there are objectives covered by the test.

The area of reliability needs much more research and discussion before an index is accepted in the field.

Item analysis. In test item construction of a NRM, a test writer wants variability, so questions that are too hard or too easy are discarded. The CRM test writer is mainly concerned with making sure that the test items accurately sample the range of criterion behavior being measured. The items must possess congruency with the class of eligible behaviors as prescribed by an instructional objective. The items can be difficult or easy, discriminating or indiscriminating, but must reflect the domain of relevant tasks. (Popham, 1971)

After the items have been formed into a test and results received from administering it, there is the procedure of analyzing and improving it. With NRM, item analysis techniques have been used to identify those items that were not properly discriminating among individuals. (The discriminating power of a test item is the ability to differentiate between persons possessing much of the same criterion trait and those possessing little of the trait.) Nondiscriminating items, or those not separating the more knowledgeable from the less knowledgeable, are usually those that are too hard, too easy, and/or ambiguous. (Ahmann, 1962)

Osburn (1968) has made the following observation about traditional item analysis techniques as applied to CRM:

> It is evident that these procedures may bias the inference regarding a person's true score on the universe of content, and the nature of the bias will generally be unknown. . . . Rejection of the item always implies rejection of the class of items to which the item belongs or at least a modification of the generating rule that specifies the item class. (p. 99)

Jackson (1970) remarked that it is difficult to see how item selection could legitimately be influenced by item analysis data because the comparability of test scores and behavioral standards are postuated upon a systematic sampling of tasks from a universe of content.

However, other people say that there is some value in item analysis techniques. Popham and Husek (1969) have suggested that a nondiscriminating item should remain on the test if the item reflects an important attribute of the criterion.

Gronlund (1965) reaffirmed this by writing:

> . . . a low index of discriminating power should alert us to the possible presence of technical defects in a test item but it should not cause us to discard an otherwise worthwhile item. A well-constructed achievement test will, of necessity, contain items with low discriminating power and to discard them would result in a test which is less, rather than more, valid. (p. 214)

Popham and Husek (1969) proposed that a positively discriminating item is a good quality to have on a CRM, and naturally should be kept. However, a negatively discriminating item, one which is answered correctly more often by the less knowledgeable than by the more knowledgeable, should be treated in the same way on both types of measures. It should be revised or thrown out.

Cox and Vargas (1966) investigated two indexes of a CRM item's ability to discriminate between pre- and post-instructional performance. One index was computed by subtracting the percentage of individuals who passed the item on the

pretest from the percentage who passed it on the posttest. The other method was the common upper group minus lower group discrimination index. The researchers concluded that the traditional way (upper minus lower group) could not be used but that the pretest-posttest method should warrant consideration when using CRM.

Jackson (1970) stated that two groups of people could be used with Cox and Vargas's procedure as long as one was known to have mastered the behavior domain to a greater degree than the other. It would be necessary to revise the domain under which a test was developed if certain items were non-discriminating between groups. This type of analysis could also be used as an empirical check on the validity of the hypothetical constructs that the test intended to measure.

A related concept to reliability is the length of the test. If CRM is used to evaluate a program or treatment, the same tests (or an equivalent form) need not be used. Cronbach (1963) and Husek and Sirotnik (1968) have shown that the concept of item sampling in which different people complete different items is highly appropriate in evaluating the adequacy of treatments. Thus, there could be a sampling of more behavior with shorter tests by constructing different forms to be administered to individuals in the treatment group.

In summary, traditional item analysis methods can be used with CRM's, to a certain extent, but it must be remembered

that discrimination is only a warning flag. Even if an item is negatively discriminating, it may be caused by an instructional deficiency or the presence of ambiguity, clues, and other technical defects in the item. (Gronlund, 1965) More developmental work on item analysis procedures, especially when only one test administration is possible, is needed.

Reporting and Interpretation. Flanagan (1951) has said that ". . . test scores are meaningful and valuable to the extent that they can be interpreted in terms of capacities, abilities, and accomplishments of educational significance." (p. 695) Ebel (1962) has pointed out that something important tends to get lost when raw scores are transformed into standard scores. "What gets lost is a meaningful relation between the scores on the test and the character of the performance it is supposed to measure." (p. 17) Ebel has advocated the use of "content standard" test scores by building meaning into the test, and hence into the test score, by a systematic, explicitly specified process of test construction.

Both NRM and CRM aid in making decisions about individuals and training treatments. The methods of norm-referenced reporting are through group-relative descriptions such as percentile ranking and standard scores. Thus, by a single score, it is possible to tell how well an individual performed in relation to the group. (Seashore, 1955)

When interpreting an individual's performance on a CRM, group-relative indices are not appropriate. Criterion

tests yield scores which are essentially "on-off" in nature. That is, the student has either mastered the criterion or has not. (Popham and Husek, 1969) In practice, however, a range of acceptable performance exists so several on-off scores should be established. (Garvin, 1971)

If an instructional objective of a carpentry training program was to be able to identify different types of hand tools used in carpentry, a 20-item objective test could be constructed and a required proficiency level set. The experts may set the minimum proficiency level at 90 percent, thus allowing error on 2 of the 20 items. In reporting an individual's performance, one alternative is that the person has reached the minimum cut-off score (90 percent) or has not. If the level is not met, the individual could not move on to the next topic, and remedial instruction would be needed. (Popham and Husek, 1969)

To report the degree of less than criterion level depends on the use of the test scores. If, for example, there are two kinds of remedial programs available, one for those close to criterion, and one for those far from criterion, the degree of performacne would be appropriate to report. (Popham and Husek, 1969)

Using CRM, the number of individuals who achieved the pre-established criterion level could be reported. Although this seems to be little data, it tells exactly the proportion of learners who did not achieve the criterion level.

The traditional descriptive statistics, such as means and standard deviations, could still be used since it is necessary to know the average performance produced by the treatment in addition to its variance. (Popham and Husek, 1969)

Millman (1972) advocated a new grading system based on CRM:

> When criterion-referenced measurement is used to guide and monitor the instructional program, it is a logical next step to have the learner's grades consist of check marks opposite instructional objectives which indicate which skills and understandings have been acquired. (p. 280)

The examples of Job Corps Training Achievement Records (1974), found in Appendix A, are similar to what Millman has called for.

For a more complete and theoretical discussion of developing and analyzing CRM's, see Swezey, Pearlstein, and Ton (1974), and Swezey and Pearlstein (1974).

The Application of Criterion-Referenced Measurement

There have been many related areas and spin-offs from the CRM movement, including mastery learning (Block, 1971), domain-referenced testing(Hively, 1968, 1974), objective-referenced testing (Baker, 1972), performance testing (Osborn, 1974), and competency-based education (Burns and Klingstedt, 1972).

Prager et al. (1972) designed a CRM program called Individual Achievement Monitoring System (IAMS) for use with handicapped people. Popham (1973) used CRM in teacher performance testing. In the Experimental Volunteer Army

Training Program, Taylor, Michael, and Brennan (1973) used performance tests for different military occupations.

An instructional innovation that has incorporated the use of CRM has been the systems approach to curriculum design. The following section illustrates how Butler (1972) has proposed CRM should be utilized in designing vocational and technical training programs.

Butler's System

Butler (1972), a vocational educator and currently director of curriculum research and development at the New England Resource Center for Occupational Education, has developed the training systems model shown in Figure 1.

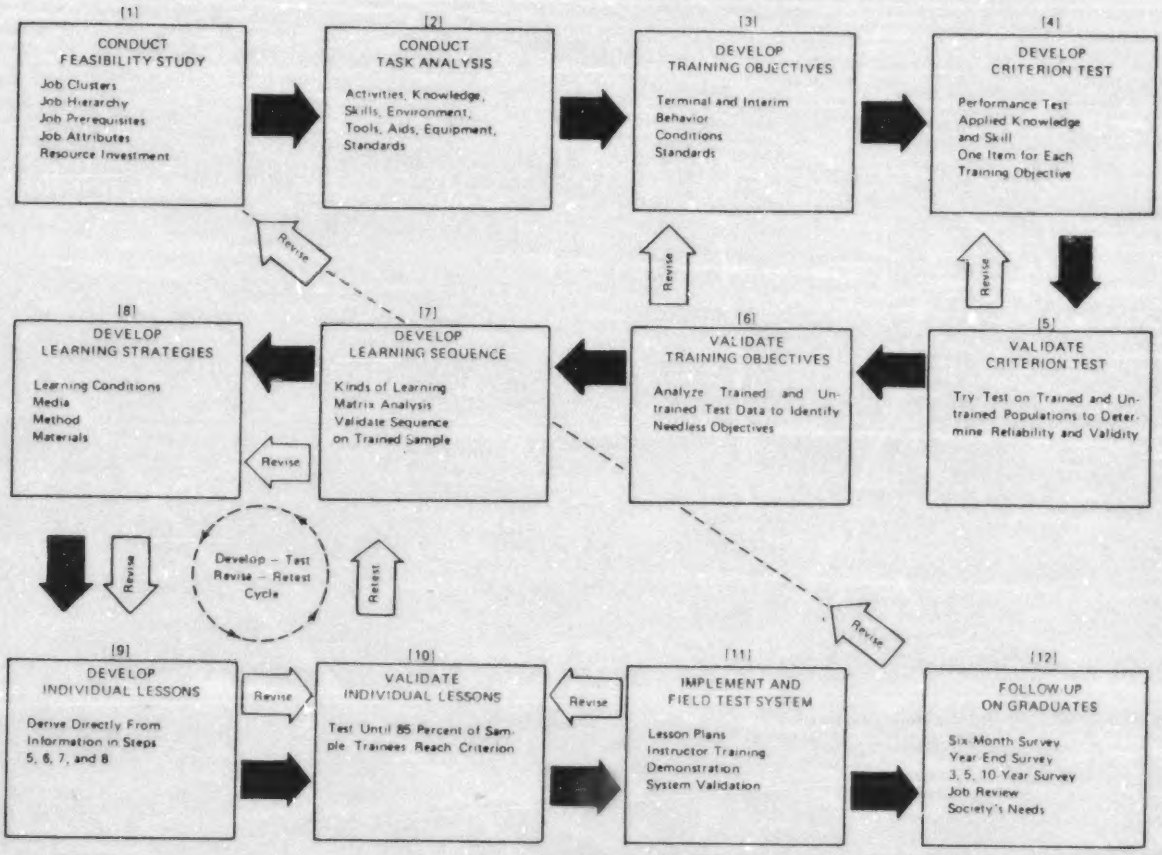

Figure 1. Butler's Training Systems Approach. (From Butler, 1972, p. 53)

Conduct feasibility study. The first step in Butler's system is an analysis of trends with regard to job markets and occupational patterns; trends in economic, business, agricultural, and industrial expansion; types of jobs and worker competencies needed; availability of training programs and facilities, and their costs; and other related information.

Conduct task analysis. After the decision has been made that a specific training program or course is needed, a job/task analysis is conducted. The job/task analysis is the foundation upon which the training objectives, content, sequence, methods, media, and evaluation are based. The job/task analysis is a summary of the behavioral content of a job broken down into duties, tasks, activities, and actions. Each task, which is "a logical and necessary step in the performance of a duty" (p. 74), should be described in the following terms:

- The cues, signals, and indication that call for the action or reaction.
- The control, object, or tool to be used or manipulated.
- The action or manipulation to be made.
- The cues, signals, and indications (feedback) that the action taken is, or is not, correct and adequate. (p. 75)

Working conditions, tools and equipment, and standards of performance are necessary for each task.

There are many possible sources of information to consult in writing a job/task analysis, such as training

literature, manuals, textbooks, The Dictionary of Occupational Titles, professional associations, trade unions, and governmental agencies. However, the most reliable and valid source is the incumbent worker. Morsch (1964) discussed seven methods of job analysis which could be used: the questionnaire-survey, individual interview, observation interview, group interview, daily diary method, work participation method, and critical incident technique.

Butler (1972) stressed that more " . . . emphasis should be placed on observation and interview of the apprentice or entry-level worker to find out what he actually does on the job . . . ". (p. 78)

Develop training objectives. Based on the task analysis, the designer must derive explicit statements about what a student, upon completion of the training program, will be able to do. Training objectives must be described in observable and measurable terms. Butler uses Mager's (1962) formula for writing objectives, whereby the conditions and limitations, overt behavior displayed by the student, and performance standards must be specified. Both terminal (unit, course, program) objectives and interim or enabling (lesson, activity, module) objectives must be specified. These may be directly coupled to broad goal statements and possibly even broader educational or philosophical constructs.

Develop criterion tests. Criterion tests are used in the early stages of design to determine validity of the

objectives, and later to provide feedback and help perform summative evaluations of the entire course or training program.

Validate the criterion tests. In order to validate the criterion test it is administered to an untrained-unskilled group and to a trained-skilled group and a correlation is computed to obtain validity and reliability coefficients. Test item analysis at this point calls for interpretations similar to the following: (a) if, for a given test item, the majority of untrained group responses are correct, the item has little or no validity or reliability; and conversely, (b) if, for a given test item, the majority of trained group responses are incorrect, the item likewise has little or no validity or reliability.

Validate training objectives. The criterion test should contain at least one item for each objective, but no more than five items for each objective, otherwise the test becomes too long for practical purposes. Validating the criterion test and validating training objectives can be accomplished concurrently, provided the test item itself is not at fault. Interpretations similar to those made in the preceding step are employed in this step; e.g., (1) if, for a given test item and its companion objective, the majority of untrained group responses are correct, there may be no need to include that objective in the curriculum; and, (2) if, for a given test item and its companion objective, the

majority of trained group responses are incorrect, there may be no need to include that objective in the course because, apparently, the worker can perform on the job without that knowledge or skill. According to Butler's model, the initial design phase has been completed at this point, but the remaining phases also require validation considerations.

Develop learning sequence. The determination of the learning sequence is done according to the duties, tasks, and activities provided in the job/task analysis. The following chart shows a pyramidal form of learning structure and sequence.

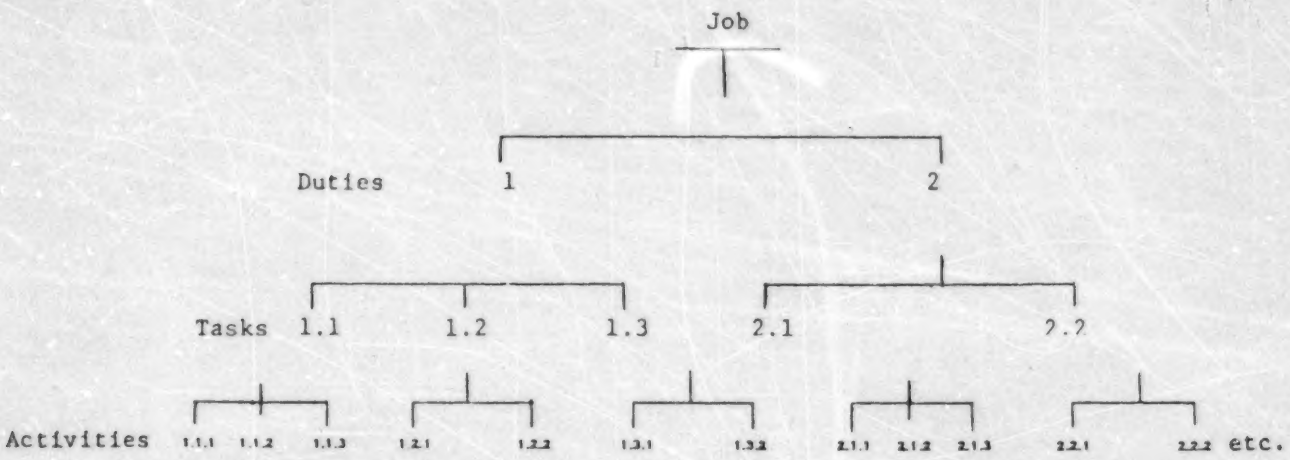

Figure 2. Pyramidal Form of Learning Structure and Sequence. (From Butler, 1972, p. 114)

Activities, tasks, and duties are structured (and learned) in both a vertical and horizontal sequence. The learning of one is dependent upon accomplishment of those which precede it. Most curriculum experts recognize that sequencing must

be approached with a great deal of flexibility. The general guideline of efficiency should influence sequencing.

Butler set forth a matrix analysis technique for preparing the course outline in which supporting knowledges and skills for activities, tasks, and duties are listed. The learning sequence can be plotted by starting with the terminal objective and working backward through each preceding prerequisite--in essence, from the complex back to the simple. Butler suggested listing all terms, concepts, rules, and principles which pertain to each objective. Each number is then placed in a two-dimensional matrix (discrimination-association) along a diagonal line from top left to bottom right. Associations then are marked in the common squares above the diagonal, and discriminations are marked in the common squares below the diagonal. By shuffling and reshuffling, a rearranged matrix can be plotted which depicts an optimum clustering of discriminations and associations around the diagonal, which results in the best sequencing. The clusters tend to depict broad concepts in the curriculum.

Validating the sequence also is accomplished with the criterion test which has been validated and revised. The test is given to a group of trained individuals, i.e., as a post-test to persons who just completed the program, or to those who have been on the job about six months. In the analysis of these scores, one looks for the dependency and interdependency between and among units, lessons, or fairly

large blocks of curriculum content.

Butler indicated that the test data should be analyzed with two basic questions in mind: (1) Did the majority of those students who correctly performed a subordinate unit also correctly perform the following and supposedly dependent unit?; and, (2) Did the majority of those who correctly performed the higher unit also perform the subordinate unit correctly? If , for a tested trained sample, the answers to both questions are affirmative, then the sequence is valid. If, for only 85% of the sample, the answers are affirmative, then the sequence is probably valid. The following chart provides a summary for analyzing criterion test data from a sample trained population.

Trained Sample (only correct performance)	Performance	Implications
Performs unit (100%)	85% perform sub-unit	Possible correct sequence
Performs sub-unit (100%)	85% perform unit	Possible correct sequence
		Taken together, a certainty the sequence is correct
Performs unit (100%)	85% perform sub-unit	Possible correct sequence
Performs sub-unit (100%)	50% fail to perform unit	Possible incorrect sequence
		Taken together, indicates bad test item
Performs unit (100%)	50% fail to perform sub-unit	Possible incorrect sequence
Performs sub-unit (100%)	85% perform unit	Possible correct sequence
		Taken together, indicates bad test item
Performs unit (100%)	50% fail to perform sub-unit	Possible incorrect sequence
Performs sub-unit (100%)	50% fail to perform unit	Possible incorrect sequence
		Taken together, a certainty the sequence is incorrect

Table 3. Validating Content Sequence. (From Butler, 1972, p. 125)

The foregoing procedure is used on a pair of tasks in a hierarchy. Suppose the hierarchy consisted of three or more tasks and validation is still required. Recent research has gone in the direction of trying to discover such hierarchies and their properties, and validation procedures are under study, using factor analysis techniques. The reader may wish to refer to "A Method for Validating Sequential Instructional Hierarchies," by P. W. Airasian, in the December, 1971 issue of Educational Technology. Airasian's method is based on calculation of conditional item difficulty indices and facilitates the pinpointing of sequential levels within a hierarchy which require revision.

Develop learning strategies. There are no feasible validation procedures for developing learning strategies which are not costly and time consuming to use. Media are selected according to those that will do an effective job for the least cost. Combinations of the different media usually should be considered.

Validation is influenced by the media. Test scores may be low for students with reading problems, but the same test scores may be improved by using audio media instead of printed media. The objectives and student learning styles are the prime determinants in developing the learning strategies.

Develop instructional lessons. This is the point

where a test model of the instructional system is produced. Two documents are needed: (1) the system development plan, and (2) the instructor's manual or guide.

The system development plan contains: (1) task analysis summary forms; (2) validated objectives in validated sequence, supported by a summary of the validation data; (3) validated criterion test items in validated sequence, supported by a summary of the validation data; (4) outline of instructional strategies with associated content (objectives) identified; and (5) production and testing plans for the system.

The design and format of the individual learning units may vary greatly, but each should contain the following: (1) the performance objectives; (2) the knowledges and skills to be gained; (3) a list of tools, equipment, supplies, references, etc.,needed for the unit; (4) a learning activity guide; (5) interim progress checks and student self-evaluations; and (6) an instrument to serve as a pre-test and/or a post-test for evaluations by the instructor.

Validate individual lessons. At this point, each unit is tested and revised until 85% of sample trainees reach the criterion.

Revision may require resequencing and adoption of new learning strategies. Initial testing is done on an individual or one-to-one basis, with two or three sample trainees who have upper-level ability. Minor revisions may

be made at this point; however, if major revision is indicated, two or three more individual tryouts should be conducted.

Small-group tryout is then conducted on 6 to 10 students who represent the range of ability and background of the target population. Criterion test data are again used to locate trouble spots and revision is made. At this point, 85% of the students should be performing correctly on the criterion test.

Final tryout is made on a large group of 30 to 50 students under conditions which approximate actual training. This tryout is conducted by the curriculum designer along with the instructor. A group this size is needed to verify or validate previous design results. Final revision is made following this tryout.

Implement and field test system. This is done under actual classroom conditions. The instructor's role in the instructional system is explicated at this point, and an instructor's manual is developed. The teacher becomes a manager and facilitator of learning and his tasks are as follows: (1) diagnose individual learning needs; (2) prescribe learning experiences; (3) provide proper materials and equipment at right time; (4) test and evaluate individual progress; (5) compile individual and group progress records; (6) provide tutorial and counseling help: (7) provide motivational reinforcement; (8) provide supplementary

materials and experiences; (9) coordinate individual, small-group, and large-group learning activities; (10) coordinate use of learning materials and equipment; and (11) evaluate feedback data on effectiveness of learning.

The instructor's manual should contain: (1) course description; (2) student population description; (3) performance objectives; (4) criterion tests; (5) system performance data; and (6) suggestions for administering the system.

Field testing is the final phase of the systems development process. This means the program is monitored, evaluated, and subsequently revised continuously for as long as it is in use. This phase may be more appropriately referred to as system "institutionalization." Constant monitoring and analysis of criterion test data will continue to point the way for needed revision.

Butler pointed out that a training system is never a finished product but rather it is constantly in process.

Follow-up on graduates. Effective guidance and placement are important in a systems approach. Longitudinal planning for follow-up at 1-year, 3-year, 5-year, or 10-year intervals should be started. Follow-up to obtain details of occupational patterns, changes in needed competencies, job adjustment problems, and work satisfaction indices, all can be used as feedback to improve the instructional system.

Chapter 3

SUMMARY AND CONCLUSIONS

CRM, in general, is the assessment of an individual's performance based on the degree to which his or her behavioral responses resemble the desired performance or criterion at a specified level. The individual's score is directly interpretable in terms of a specified universe of content and instructionally relevant tasks.

Both NRM and CRM help in making basic decisions concerning individuals and programs. However, the score interpretation is different in these two measures. A normative score indicates how well an individual performed on a measure in relationship to others on the same measure. A criterion score is directly interpretable as to what an individual can or cannot do in relationship to a specified universe of content.

The major differences between the two measures lie in the purpose of the test, the manner in which it is constructed, the specificity of the information obtained about the domain of relevant tasks, the generalization of the test performance, and the use of the score.

In determining which type of measurement to use, if there is the need for selectivity or competitive comparison among individuals, NRM should be used. CRM should be used

to determine whether a person has mastered certain knowledges, understandings, and skills. CRM can also be used with any type of programmed learning or individualized instruction, and for promotion and licensing procedures.

When writing a CRM, the test constructor must make sure that the test items accurately sample the range of criterion behaviors being measured. Criterion relevance, deficiency, and contamination should be analyzed. The items must possess congruency with the universe of instructionally relevant tasks.

The first step in evaluating training outcomes is to define precisely what is to be measured. This is accomplished by writing behavioral or performance objectives for all desired outcomes. These behavioral objectives must be translated into specific test tasks which form the basis for inference that the behaviors have been acquired by the individual.

The most important requirement when writing a CRM is that an objective, systematic procedure be used to specify the domain of tasks required to be performed. One such method is through the use of an item form, which consist of a general form and generation rules which specifically defines the required tasks. The item form can be used to generate many different items with a fixed syntactical structure. Thus, a collection of item forms define the universe of content for the test.

The major concern of CRM experts is the need for evaluating how "good" a criterion-referenced test is. While there are many textbooks and articles writtern about the well-honed mental test theory procedures (norm-referenced tests), there are very few guides available on criterion-referenced tests. Since most of the traditional theories and formulas for determining the adequacy of a NRM are based on variance, they cannot be applied to criterion-referenced measures. Variability is irrelevant with CRM because the meaning of the scores flows directly from the connection between the items to the criterion.

Several variations of traditional test theory have been suggested for evaluating the adequacy of a criterion measure. Content validity is the main method to evaluate if the test measures what it purports to measure. Equivalent forms and sequentially scaled tests have been proposed to be used to estimate the consistency or reliability of the test. A pretest, posttest discrimination index could be used to evaluate a test item, and the traditional upper group minus lower group could be used with limitations.

Conclusions

The literature would appear to support the following conclusions:

1. Although experts do not agree on a single definition of criterion-referenced measurement, all variations have in common an emphasis on the interpretation of the

individual's score, which represents what an individual can do relative to the instructional objectives of a program.

2. Criterion-referenced information is valuable in making instructional decisions based on what a person can do at a certain time in the training cycle. If training is going to become more adaptive to the individual, this input is a necessity.

3. CRM have focused much attention on behavioral objectives and desired trainee outcomes. Detailed specifications of test construction processess and experimental evidence relating behavior to test performance appear to be a promising approach to the measurement of competencies in training.

4. Behavioral objectives must be carefully written in order to more validly direct the instructional design and measure its effectiveness.

5. More than one method should be used to validate any desired CRM in order to decrease the error that is associated with its measurement.

6. It is difficult to develop objective procedures necessary for CRM of complex behavior. For complex behavioral domains, until explicit models stated in measurable terms are developed, there is too much of a degree of subjectivity in this type of test construction.

7. CRM supplements but should not replace normative tests in training. Both are essential for making decisions

about the training process. The more simple, clear, and direct test results can be presented, the more useful and instructionally fruitful tests are likely to be.

8. CRM seem interesting and relevant for today's training systems, but there is need for research, both theoretically and empirically, before extensive use of it can be recommended in an instructional environment.

REFERENCES

Airasian, P. W., and G. F. Madans. "Criterion-Referenced Testing in the Classroom," Measurement in Education, May, 1972, 1-8.

Astin, A. W. "Criterion-Centered Research," Educational and Psychological Measurement, 1964, 24, 807-822.

Ayres, L. P. A Scale for Measuring Quality of Handwriting of School Children. New York: Russell Sage Foundation, 1912.

Baker, E. L. "Teaching Performance Tests of Dependent Measures in Instructional Research." Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

Block, J. H. (Ed.) Mastery Learning: Theory and Practice. New York: Holt, Rinehart, and Winston, 1971.

Boehm, A. E. "Criterion-Referenced Assessment for the Teacher," Teachers College Record, September 1973, 117-126.

Bormuth, J. R. On the Theory of Achievement Test Items. Chicago: University of Chicago Press, 1970.

Burns, R. W. and J. L. Klingstedt. (Eds.) Competency-Based Education: An Introduction. Englewood Cliffs, N. J.: Educational Technology, 1973.

Cattell, R. B. , and H. J. Butcher. The Prediction of Achievement and Creativity. New York: Bobbs-Merrill, 1968.

Coulson, J. E., and J. F. Cogswell. "Effects of Individualized Instruction on Testing," Journal of Educational Measurement, Spring, 1965, 59-64.

Cox, R. C. "Evaluative Aspects of Criterion-Referenced Measures," Criterion-Referenced Measurement, An Introduction, Ed. W. James Popham. Englewood Cliffs, N. J.: Educational Technology Publications, 1971, 67-75.

_____, and J. S. Vargas. "A Comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests." Paper read at the Annual Meeting of the National Council on Measurement in Education, February, 1966, Chicago, Illinois.

Cronback, L. J. "Test Validation," Educational Measurement. 2nd ed., Ed. R. L. Thorndike. Washington, D. C.: American Council on Education, 1971, 443-507.

_____. "Validity," Encyclopedia of Educational Research, Ed. C. W. Harris. New York: Macmillan, 1960.

Cureton, E. E. "Validity," Educational Measurement, Ed. E.F. Lindquist. New York: Macmillan, 1951, 621-694.

Ebel, R. L. "Must All Tests Be Valid?" American Psychologist, 1961, 16, 640-647.

_____. "Content Standard Test Scores," Educational and Psychological Measurement, 1962, 22, 15-25.

_____. Measuring Educational Achievement. Englewood Cliffs, N. J.: Prentice-Hall, 1965.

_____. "Content Standard Test Scores: Limitations," School Review, 1971, 79, 282-288.

_____. "Some Limitations of Criterion-Referenced Measurement." American Educational Research Association. ERIC Microfilms, 1970, ED 038 670.

Flanagan, J. C. "Units, Scores, and Norms," Educational Measurement, Ed. E. F. Lindquist. Washington, D. C.: American Council on Education, 1950, 695-763.

Fremer, J. "Criterion-Referenced Interpretations of Survey Achievement Tests." Princeton, N. J.: Educational Testing Service, 1972. (Mimeographed.)

Gagne, R. M. The Conditions of Learning. New York: Holt, Rinehart, and Winston, 1965.

Garry, R. The Psychology of Learning. Washington, D. C.: The Center for Applied Research in Education, 1963.

Garvin, A. D. "The Applicability of Criterion-Referenced Measurement by Content Area and Level," Criterion-Referenced Measurement, Ed. W. J. Popham. Englewood Cliffs, N. J.: Educational Technology, 1971, 55-63.

Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, 1963, 18, 519-521.

_____, and R. C. Cox. "Criterion-Referenced Testing for the Measurement of Educational Outcomes," Instructional Process and Media Innovation, Ed. R. A. Weisgerber. Chicago: Rand-McNally, 1968, 545-550

_____, and A J. Nitko. "Measurement in Learning and Instruction," Educational Measurement. 2d ed., Ed. R. L. Thorndike. Washington, D. C.: American Council on Education, 1971, 625-670.

Goldstein, I. L. Training: Program Development and Evaluation. Monterey, California: Brooks/Cole, 1974.

Gronlund, N. E. Measurement and Evaluation in Teaching. 2d ed. New York: Macmillan, 1971.

_____(Ed.). Readings in Measurement and Evaluation. New York: Macmillan, 1968.

Harris, C. W. "An Interpretation of Livingston's Reliability Coefficient for Criterion-Referenced Tests." Journal of Educational Measurement, 1972, 9, 27-29.

Hively, W. "Introduction to Domain-Referenced Testing," Educational Technology, June, 1974, 5-9.

_____, et al. "A 'Universe-Defined' System of Arithmetic Achievement Tests," Journal of Educational Measurement, Winter, 1968, 275-295.

Horrocks, J. E., and T. J. Schoonover. Measurement for Teachers. Columbus, Ohio: Charles E. Merrill, 1968.

Husek, T. R., and K. Sirotnik. "Item Sampling in Educational Research: An Empirical Investigation." Paper presented at the Annual Meeting of the American Educational Research Association, February, 1968, Chicago, Illinois.

Jackson, R. "Developing Criterion-Referenced Tests." Princeton, N. J.: Educational Testing Service, June, 1970. (Mimeographed.)

Lindquist, E. F. The Impact of Machines on Educational Measurement. Bloomington, Ind.: Phi Delta Kappa International, 1968.

Livingston, S. A. "A Classical Test-Theory Approach to Criterion-Referenced Tests." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972. (a)

_____. "A Reply to Harris' An Interpretation of Livingston's Reliability Coefficient for Criterion-Referenced Tests." Journal of Educational Measurement, 1972, 9, 3. (b)

Lennon, R. T. "Assumptions Underlying the Use of Content Validity," Educational and Psychological Measurement, May, 1956, 297-304.

Lord, F. M., and M. R. Novick. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Lynton, R. P. and V. Pareek. Training for Development. Homewood, Illinois: Dorsey Press, 1967.

Mager, R. F. Preparing Instructional Objectives. Belmont, Calif.: Fearon, 1962.

Mehrens, W. A., and I. J. Lehmann. Standardized Tests in Education. New York: Holt, Rinehart, and Winston, 1969.

Meredith, K. E. , and D. L. Sabers. "Using Item Data for Evaluating Criterion-Referenced Measures With an Empirical Investigation of Index Consistency." Paper presented at the Annual Meeting of the Rocky Mountain Psychological Association, Albuquerque, 1972.

Millman, J. "Criterion-Referenced Measurement: An Alternative," The Reading Teacher, December, 1972, 278-281.

Mirsberger, G. E. "The Four Crucial Phases of Evaluation," Training, August, 1974, 34-35.

Oakland, T. "An Evaluation of Available Models for Estimating the Reliability and Validity of Criterion-Referenced Measures." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Osburn, H. G. "Item Sampling for Achievement Testing," Educational and Psychological Measurement, 1968, 95-104.

Osborn, W. C. "Framework for Performance Testing," Training in Business and Industry, May, 1974, 28-31.

Popham, W. J. (Ed.). "Indicies of Adequacy for Criterion-Referenced Test Items," Criterion-Referenced Measurement, An Introduction. Englewood Cliffs, N. J.: Educational Technology, 1971.

_____. "Applications of Teaching Performance Tests to Inservice and Preservice Teacher Training." Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, 1973.

_____, and T. K. Husek. "Implications of Criterion-Referenced Measurement," The Journal of Educational Measurement, Spring, 1969, 1-9.

Prager, B. B., et al. "Adapting Criterion-Referenced Measurement to Individualization of Instruction for Handicapped Children: Some Issues and a First Attempt." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 1972.

Seashore, H. G. "Methods of Expressing Test Scores," Test Service Bulletin No. 48. New York: The Psychological Corp., 1955.

Shoemaker, D. M., and H. G. Osburn. "Computer-Aided Item Sampling for Achievement Testing," Educational and Psychological Measurement, 1969, 29, 165-172.

Swaminathan, H., et al. "Reliability of Criterion-Referenced Tests: A Decision-Theoretic Formulation." Journal of Educational Measurement, 1974, 11, 263-268.

Swezey, R. W. and R. B. Pearlstein. Developing Criterion-Referenced Tests. Reston, Va.: Applied Science Associates, 1974.

_____, _____, and W. H. Ton. Criterion-Referenced Testing: A Discussion of Theory and of Practice in the Army. Reston, Va.: Applied Science Associates, 1974.

Taylor, J. E., et al. The Concepts of Performance Oriented Instruction Used in Developing the Experimental Volunteer Army Training Program. HumRRO Technical Report TR-73-3, 1973.

Thorndike, E. L. "Handwriting," Teachers College Record, March, 1910, 83-175.

_____. "The Nature, Purposes and General Methods of Educational Products," The Measurement of Educational Products. Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School, 1918.

_____. "The Original Nature of Man, Educational Psychology. Vol I. New York: Teacher's College, Columbia University, 1913.

Thorndike, R.L. & E. Hagen. Measurement and Evaluation in Psychology and Education. New York: John Wiley and Sons, 1955.

Travers, S. "Behavioral Objectives: An Overview." A research paper presented to PSYC 731, Training Procedures and Evaluation in Organizational Settings, University of Maryland, February, 1975.

Trow, W. C. "Foreword," Criterion-Referenced Measurement, An Introduction, Ed. W. J. Popham. Englewood Cliffs, N. J.: Educational Technology, 1971.

Tuckman, B. W. Conducting Educational Research. New York: Harcourt Brace Jovanovich, 1972.

Tyler, R. W. Constructing Achievement Tests. Columbus, Ohio: Ohio State University, 1934.

Wang, M. C. "Approaches to the Validation of Learning Hierarchies." Western Regional Conference on Testing Problems (Proceedings) 1969. Princeton, N. J.: Educational Testing Service, 1969, 14-38.

ADDENDUM

Butler, F.C. Instructional Systems Development for Vocational and Technical Training. Englewood Cliffs, N.J.: Educational Technology, 1972.

Cartier, Francis A. "Criterion-Referenced Testing of Language Skills." ERIC Microfilms, 1968, ED 020 515.

Cox, R. C., and G. T. Graham. "The Development of a Sequentially Scaled Achievement Test," Journal of Educational Measurement, Fall 1966, 147-150.

Cronback, L. J., and G. C. Gleser. Psychological Tests and Personal Decision. 2d ed. Urbana, Ill.: University of Illinois Press, 1965.

Day, G. F. "An Investigation into the Use of Instructional Systems Approaches in Training and Education." Unpublished paper, University of Maryland, 1975.

Edgerton, H. A. "The Place of Measuring Instruments in Guidance," The Measurement of Student Adjustment and Achievement, Ed. T. Donahue, and others. Ann Arbor, Mich.: University of Michigan Press, 1949, 51-58.

Ferguson, R. L. "Computer-Assisted Criterion-Referenced Testing," Working Paper No. 49. Pittsburgh University, Pennsylvania. Learning Research and Development Center. ERIC Microfilms, 1969, ED 040 061.

Green, J. A. Introduction to Measurement and Evaluation. New York: Dodd, Mead and Co., 1970.

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

Job Corps. Training Achievement Records. Washington, D. C.: U. S. Government Printing Office, 1974.

Karmel, L. J. Measurement and Evaluation in the Schools. New York: Macmillan, 1970.

Lindquist, E. F. A First Course in Statistics. Boston: Houghton Mifflin, 1942.

Lord, F. M., and M. R. Novick. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Nagle, B. F. "Criterion Development," Personnel Psychology, 1953, 6, 271-288.

Simon, G. B. "Comments on 'Implications of Criterion-Referenced Measurement'," Journal of Educational Measurement, Winter, 1969, 259-260.

Thorndike, R L. Personnel Selection. New York: Wiley, 1949.

Thornton, R. F., and R. G. Wasdyke. "The Development of Criterion-Referenced Tests Using the Behavioral Consistency Model." Princeton, N. J.: Educational Testing Service, 1972. (Mimeographed.)

Appendix A

TRAINING ACHIEVEMENT RECORD

Name _____ SSN _____ Date Trainee Entered Training _____

Title SECRETARY - STENOGRAPHER DOT Code 201.368 Certified by _____ Form No. 440.185 A

Achieved Individual Marketable Skill		PERFORMANCE				KNOWLEDGE			
		1	2	3	4	a	b	c	d
Safety									
1. Use safe practices in operating equipment & carrying out all functions	4d								
2. Use appropriate safety procedures when using vaporized/caustic products	4d								
General Skills									
3. Know and use correct oral & written English	3c								
4. Use directories & other reference materials correctly	3c								
5. Know & use office forms, supplies & equipment	3c								
6. Maintain appearance of office	3c								
Correspondence Skills									
7. Take dictation in shorthand to a minimum of 90 wpm with 97% accuracy on a three minute timing	4d								
8. Transcribe dictated materials to standards of mailability to a minimum 25 wpm	4d								
9. Type at least 50 wpm with 2 errors on an electric typewriter during a 5 minute timing	4d								
10. Compose written communications on own initiative or from oral instructions	3c								
11. Type from longhand notes	3c								
12. Transcribe from dictating machine	4d								
13. Type & make corrections on multiple carbon copies	3c								
14. Type all styles letters, memos, reports & stencils using appropriate correction material	3c								
15. Correct in transcription grammar, arithmetic, facts, etc	3c								
Public Relations									
16. Answer telephone & give information & route calls	3c								
17. Place both long-distance & local outgoing calls	3c								
18. Ascertain visitor's business & conduct them appropriately	3c								
19. Greet visitors & provide for their comfort	3c								
Administerative Skills									
20. Schedule appointments for employer & maintain the calendar	3c								
21. Read & route incoming mail	3c								
22. Organize material for correspondence to be answered by employer	3c								
23. Arrange itinerary, reservations & materials for travel	3c								
24. Handle meeting & conference arrangements	3c								
25. Supervise or coordinate w/other clerical workers	3c								
26. Organize & maintain files; retaining confidential files; set up retention schedules for files	3c								
27. Follow through on projects to see that deadlines are met	3c								
28. Recommend new clerical services, systems or equipment	3c								
29. Organize & schedule work on a daily & long range basis	3c								
Clerical Skills									
30. Maintain telephone & address records	3c								

31. Keep petty cash funds & bank account records	3c								
32. Know & use telegraphic services	3c								
33. Operate office machines including duplicating, copy, calculating, etc.	3c								
23. Record minutes of staff meetings & conferences	3c								
35. Purchase & receive supplies & equipment	3c								
Additional Related Training Elements									

EDUCATION, TECHNICAL KNOWLEDGE--JOB PHYSICAL PROFILE		PERFORMANCE 1	2	3	4	KNOWLEDGE a	b	c	d
1. Use instructions furnished in written, oral, diagram or schedule form	4d								
2. Use arithmetic; apply practical algebra and geometry	4d								
3. Read and interpret technical materials	4d								
4. Prepare reports and summaries, conforming to good English usage	4d								
1. Able to lift 10 lbs. maximum; sit most of the day	4								
2. Able to use fingers, hands, arms, to reach, handle, feel	4								
3. Able to speak & hear	4								
4. Able to see efficiently	4								
5. Work indoors	4								

ATTITUDES AND PROFESSIONAL ETHICS		ATTITUDES & ETHICS I	II	III	IV	V
1. Demonstrate correct safety practices on the job	IV					
2. Maintain appropriate personal hygiene and appearance	IV					
3. Arrive on the job on time	IV					
4. Is on the job every day	IV					
5. Perform work of consistently good quality	IV					
6. Function cooperatively with fellow workers	IV					
7. Treat others courteously	IV					
8. Work with even temperament	IV					
9. Accept constructive criticism	IV					
10. Follow instructions willingly	IV					
11. Deal well with supervision	IV					
12. Willingly work unusual schedules when required	IV					
13. Handle proprietary information discreetly; respect confidences	IV					
14. Respect worth of equipment, company and personal property						